IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2019

Project-Team VALDA

Value from Data

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

RESEARCH CENTER
Paris

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team VALDA

Creation of the Team: 2016 December 01, updated into Project-Team: 2018 January 01

Keywords:

Computer Science and Digital Science:

A3.1. - Data
A3.1.1. - Modeling, representation
A3.1.2. - Data management, quering and storage
A3.1.3. - Distributed data
A3.1.4. - Uncertain data
A3.1.5. - Control access, privacy
A3.1.6. - Query optimization
A3.1.7. - Open data
A3.1.8. - Big data (production, storage, transfer)
A3.1.9. - Database
A3.1.10. - Heterogeneous data
A3.1.11. - Structured data
A3.2. - Knowledge
A3.2.1. - Knowledge bases
A3.2.2. - Knowledge extraction, cleaning
A3.2.3. - Inference
A3.2.4. - Semantic Web
A3.2.5. - Ontologies
A3.2.6. - Linked data
A3.3.2. - Data mining
A3.4.3. - Reinforcement learning
A3.4.5. - Bayesian methods
A3.5.1. - Analysis of large graphs
A4.7. - Access control
A7.2. - Logic in Computer Science
A7.3. - Calculability and computability
A9.1. - Knowledge
A9.8. - Reasoning

Other Research Topics and Application Domains:

B6.3.1. - Web
B6.3.4. - Social Networks
B6.5. - Information systems
B9.5.6. - Data science
B9.6.5. - Sociology
B9.6.10. - Digital humanities
B9.7.2. - Open data
B9.9. - Ethics
B9.10. - Privacy

1. Team, Visitors, External Collaborators

Research Scientists

Serge Abiteboul [Inria, Emeritus, from Oct 2019; also ARCEP, Board Member, HDR]
Camille Bourgaux [CNRS, Researcher]
Olivier Cappé [CNRS, Senior Researcher, HDR]
Luc Segoufin [Inria, Senior Researcher, HDR]
Michaël Thomazo [Inria, Researcher]

Faculty Members

Pierre Senellart [Team leader, École normale supérieure, Professor, HDR]
Leonid Libkin [École normale supérieure, Professor, from Sep 2019]
Silviu Maniu [Univ Paris-Saclay, Associate Professor, from Sep 2019]

Post-Doctoral Fellows

Ashish Deepak Dandekar [École normale supérieure, Post-Doctoral Fellow, from Jul 2019]
Nathan Grosshans [École normale supérieure, Post-Doctoral Fellow]

PhD Students

Juliette Achddou [Numberly, PhD Student, from Mar 2019]
Julien Grange [Université Paris Diderot, PhD Student]
Karima Rafes [BorderCloud, PhD Student, until Feb 2019]
Yann Ramusat [École normale supérieure, PhD Student]
Yoan Russac [École normale supérieure, PhD Student]

Interns and Apprentices

Rémi Dupré [École normale supérieure, from Mar 2019 until Aug 2019]
Quentin Manière [Inria, from Apr 2019 until Aug 2019]

Administrative Assistant

Meriem Guemair [Inria, Administrative Assistant]

Visiting Scientist

Victor Vianu [UCSD, Université Paris Diderot, & École normale supérieure, from Jun 2018]

2. Overall Objectives

2.1. Objectives

Valda's focus is on both *foundational and systems aspects of* complex *data management*, especially *human-centric data*. The data we are interested in is typically heterogeneous, massively distributed, rapidly evolving, intensional, and often subjective, possibly erroneous, imprecise, incomplete. In this setting, Valda is in particular concerned with the optimization of complex resources such as computer time and space, communication, monetary, and privacy budgets. The goal is to extract *value from data*, beyond simple query answering.

Data management [37], [46] is now an old, well-established field, for which many scientific results and techniques have been accumulated since the sixties. Originally, most works dealt with static, homogeneous, and precise data. Later, works were devoted to heterogeneous data [35] [38], and possibly distributed [76] but at a small scale.

However, these classical techniques are poorly adapted to handle the new challenges of data management. Consider human-centric data, which is either produced by humans, e.g., emails, chats, recommendations, or produced by systems when dealing with humans, e.g., geolocation, business transactions, results of data analysis. When dealing with such data, and to accomplish any task to extract value from such data, we rapidly encounter the following facets:

- *Heterogeneity*: data may come in many different structures such as unstructured text, graphs, data streams, complex aggregates, etc., using many different schemas or ontologies.
- *Massive distribution*: data may come from a large number of autonomous sources distributed over the web, with complex access patterns.
- *Rapid evolution*: many sources may be producing data in real time, even if little of it is perhaps relevant to the specific application. Typically, recent data is of particular interest and changes have to be monitored.
- *Intensionality*[1]: in a classical database, all the data is available. In modern applications, the data is more and more available only intensionally, possibly at some cost, with the difficulty to discover which source can contribute towards a particular goal, and this with some uncertainty.
- *Confidentiality and security*: some personal data is critical and need to remain confidential. Applications manipulating personal data must take this into account and must be secure against linking.
- *Uncertainty*: modern data, and in particular human-centric data, typically includes errors, contradictions, imprecision, incompleteness, which complicates reasoning. Furthermore, the subjective nature of the data, with opinions, sentiments, or biases, also makes reasoning harder since one has, for instance, to consider different agents with distinct, possibly contradicting knowledge.

These problems have already been studied individually and have led to techniques such as *query rewriting* [59] or *distributed query optimization* [64].

Among all these aspects, intensionality is perhaps the one that has least been studied, so we pay particular attention to it. Consider a user's query, taken in a very broad sense: it may be a classical database query, some information retrieval search, a clustering or classification task, or some more advanced knowledge extraction request. Because of intensionality of data, solving such a query is a typically dynamic task: each time new data is obtained, the partial knowledge a system has of the world is revised, and query plans need to be updated, as in adaptive query processing [52] or aggregated search [75]. The system then needs to decide, based on this partial knowledge, of the best next access to perform. This is reminiscent of the central problem of reinforcement learning [73] (train an agent to accomplish a task in a partially known world based on rewards obtained) and of active learning [70] (decide which action to perform next in order to optimize a learning strategy) and we intend to explore this connection further.

Uncertainty of the data interacts with its intensionality: efforts are required to obtain more precise, more complete, sounder results, which yields a trade-off between *processing cost* and *data quality*.

Other aspects, such as heterogeneity and massive distribution, are of major importance as well. A standard data management task, such as query answering, information retrieval, or clustering, may become much more challenging when taking into account the fact that data is not available in a central location, or in a common format. We aim to take these aspects into account, to be able to apply our research to real-world applications.

2.2. The Issues

We intend to tackle hard technical issues such as query answering, data integration, data monitoring, verification of data-centric systems, truth finding, knowledge extraction, data analytics, that take a different flavor in this modern context. In particular, we are interested in designing strategies to *minimize data access cost towards a specific goal, possibly a massive data analysis task*. That cost may be in terms of communication (accessing data in distributed systems, on the Web), of computational resources (when data is produced

[1] We use the spelling *intensional*, as in mathematical logic and philosophy, to describe something that is neither available nor defined in *extension*; *intensional* is derived from *intension*, while *intentional* is derived from *intent*.

by complex tools such as information extraction, machine learning systems, or complex query processing), of monetary budget (paid-for application programming interfaces, crowdsourcing platforms), or of a privacy budget (as in the standard framework of differential privacy).

A number of data management tasks in Valda are inherently intractable. In addition to properly characterizing this intractability in terms of complexity theory, we intend to develop solutions for solving these tasks in practice, based on approximation strategies, randomized algorithms, enumeration algorithms with constant delay, or identification of restricted forms of data instances lowering the complexity of the task.

3. Research Program

3.1. Scientific Foundations

We now detail some of the scientific foundations of our research on complex data management. This is the occasion to review connections between data management, especially on complex data as is the focus of Valda, with related research areas.

3.1.1. Complexity & Logic

Data management has been connected to logic since the advent of the relational model as main representation system for real-world data, and of first-order logic as the logical core of database querying languages [37]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [63], recursive queries (Datalog), or querying of XML databases [46]. Logical formalisms facilitate reasoning about the expressiveness of a query language or about its complexity.

The main problem of interest in data management is that of query evaluation, i.e., computing the results of a query over a database. The complexity of this problem has far-reaching consequences. For example, it is because first-order logic is in the AC_0 complexity class that evaluation of SQL queries can be parallelized efficiently. It is usual [74] in data management to distinguish *data complexity*, where the query is considered to be fixed, from *combined complexity*, where both the query and the data are considered to be part of the input. Thus, though conjunctive queries, corresponding to a simple SELECT-FROM-WHERE fragment of SQL, have PTIME data complexity, they are NP-hard in combined complexity. Making this distinction is important, because data is often far larger (up to the order of terabytes) than queries (rarely more than a few hundred bytes). Beyond simple query evaluation, a central question in data management remains that of complexity; tools from algorithm analysis, and complexity theory can be used to pinpoint the tractability frontier of data management tasks.

3.1.2. Automata Theory

Automata theory and formal languages arise as important components of the study of many data management tasks: in temporal databases [36], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [42], queries are naturally given as automata; typical query and schema languages for XML databases such as XPath and XML Schema can be compiled to tree automata [67], or for more complex languages to data tree automata[4]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [50] that show that very expressive queries (from the language of monadic second-order language) can be evaluated as tree automata over *tree decompositions* of the original databases, yielding linear-time algorithms (in data complexity) for a wide variety of applications.

3.1.3. Verification

Complex data management also has connections to verification and static analysis. Besides query evaluation, a central problem in data management is that of deciding whether two queries are *equivalent* [37]. This is critical for query optimization, in order to determine if the rewriting of a query, maybe cheaper to evaluate, will return the same result as the original query. Equivalence can easily be seen to be an instance of the problem of (non-)satisfiability: $q \equiv q'$ if and only if $(q \land \neg q') \lor (\neg q \land q')$ is not satisfiable. In other words, some aspects of query optimization are static analysis issues. Verification is also a critical part of any database application where it is important to ensure that some property will never (or always) arise [48].

3.1.4. Workflows

The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of data management applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and to statically verify critical properties of the system. Different approaches to workflow specifications exist: automata-based, logic-based, or predicate-based control of function calls [34].

3.1.5. Probability & Provenance

To deal with the uncertainty attached to data, proper models need to be used (such as attaching *provenance* information to data items and viewing the whole database as being *probabilistic*) and practical methods and systems need to be developed to both reliably estimate the uncertainty in data items and properly manage provenance and uncertainty information throughout a long, complex system.

The simplest model of data uncertainty is the NULLs of SQL databases, also called Codd tables [37]. This representation system is too basic for any complex task, and has the major inconvenient of not being closed under even simple queries or updates. A solution to this has been proposed in the form of *conditional tables* [61] where every tuple is annotated with a Boolean formula over independent Boolean random events. This model has been recognized as foundational and extended in two different directions: to more expressive models of *provenance* than what Boolean functions capture, through a semiring formalism [57], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [58]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [49], [43]. Research in the past ten years has focused on a better understanding of the tractability of query answering with provenance and probabilistic annotations, in a variety of specializations of this framework [72] [62], [40].

3.1.6. Machine Learning

Statistical machine learning, and its applications to data mining and data analytics, is a major foundation of data management research. A large variety of research areas in complex data management, such as wrapper induction [68], crowdsourcing [41], focused crawling [56], or automatic database tuning [44] critically rely on machine learning techniques, such as classification [60], probabilistic models [55], or reinforcement learning [73].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [65] system result in probabilistic annotations that need to be properly modeled, stored, and queried.

Finally, complex data management also brings new twists to some classical machine learning problems. Consider for instance the area of *active learning* [70], a subfield of machine learning concerned with how to optimally use a (costly) oracle, in an interactive manner, to label training data that will be used to build a learning model, e.g., a classifier. In most of the active learning literature, the cost model is very basic (uniform or fixed-value costs), though some works [69] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [53]. These assumptions usually break when applied to complex data management problems on real-world data, such as crowdsourcing.

3.2. Research Directions

At the beginning of the Valda team, the project was to focus on the following directions:

- foundational aspects of data management, in particular related to query enumeration and reasoning on data, especially regarding security issues;
- implementation of provenance and uncertainty management, real-world applications, other aspects of uncertainty and incompleteness, in particular dynamic;
- development of personal information management systems, integration of machine learning techniques.

We believe the first two directions have been followed in a satisfactory manner. The focus on personal information management has not been kept for various organizational reasons, however, but the third axis of the project is reoriented to more general aspects of Web data management.

New permanent arrivals in the group since its creation have impacted its research directions in the following manner:

- Camille BOURGAUX and Michaël THOMAZO are both specialists of knowledge representation and formal aspects of knowledge bases, which is an expertise that did not exist in the group. They are also both interested in, and have started working on aspects related to connecting their research with database theory, and investigating aspects of uncertainty and incompleteness in their research. This will lead to more work on knowledge representation and symbolic AI aspects, while keeping the focus of Valda on foundations of data management and uncertainty.
- Olivier CAPPÉ is a specialist in statistics and machine learning, in particular multi-armed bandits and reinforcement learning. He is also interested in applications of these learning techniques to data management problems. His arrival in the group therefore complements the expertise of other researchers, and will lead to more work on machine learning issues.
- Leonid LIBKIN is a specialist of database theory, of incomplete data management, and has a line of current research on graph data management. His profile fits very well with the original orientation of the Valda project.

We intend to keep producing leading research on the foundations of data management. Generally speaking, the goal is to investigate the borders of feasibility of various tasks. For instance, what are the assumptions on data that allow for computable problems? When is it not possible at all? When can we hope for efficient query answering, when is it hopeless? This is a problem of theoretical nature which is necessary for understanding the limit of the methods and driving research towards the scenarios where positive results may be obtainable. Only when we have understood the limitation of different methods and have many examples where this is possible, we can hope to design a solid foundation that allowing for a good trade-off between what can be done (needs from the users) and what can be achieved (limitation from the system).

Similarly, we will continue our work, both foundational and practical, on various aspects of provenance and uncertainty management. One overall long-term goal is to reach a full understanding of the interactions between query evaluation or other broader data management tasks and uncertain and annotated data models. We would in particular want to go towards a full classification of tractable (typically polynomial-time) and intractable (typically NP-hard for decision problems, or #P-hard for probability evaluation) tasks, extending and connecting the query-based dichotomy [51] on probabilistic query evaluation with the instance-based one of [39], [40]. Another long-term goal is to consider more dynamic scenarios than what has been considered so far in the uncertain data management literature: when following a workflow, or when interacting with intensional data sources, how to properly represent and update uncertainty annotations that are associated with data. This is critical for many complex data management scenarios where one has to maintain a probabilistic current knowledge of the world, while obtaining new knowledge by posing queries and accessing data sources. Such intensional tasks requires minimizing jointly data uncertainty and cost to data access.

As application area, in addition to the historical focus on personal information management which is now less stressed, we target Web data (Web pages, the semantic Web, social networks, the deep Web, crowdsourcing platforms, etc.).

We aim at keeping a delicate balance between theoretical, foundational research, and systems research, including development and implementation. This is a difficult balance to find, especially since most Valda researchers have a tendency to favor theoretical work, but we believe it is also one of the strengths of the team.

4. Application Domains

4.1. Personal Information Management Systems

We recall that Valda's focus is on human-centric data, i.e., data produced by humans, explicitly or implicitly, or more generally containing information about humans. Quite naturally, we have used as a privileged application area to validate Valda's results that of personal information management systems (Pims for short) [33].

A Pims is a system that allows a user to integrate her own data, e.g., emails and other kinds of messages, calendar, contacts, web search, social network, travel information, work projects, etc. Such information is commonly spread across different services. The goal is to give back to a user the control on her information, allowing her to formulate queries such as "What kind of interaction did I have recently with Alice B.?", "Where were my last ten business trips, and who helped me plan them?". The system has to orchestrate queries to the various services (which means knowing the existence of these services, and how to interact with them), integrate information from them (which means having data models for this information and its representation in the services), e.g., align a GPS location of the user to a business address or place mentioned in an email, or an event in a calendar to some event in a Web search. This information must be accessed intensionally: for instance, costly information extraction tools should only be run on emails which seem relevant, perhaps identified by a less costly cursory analysis (this means, in turn, obtaining a cost model for access to the different services). Impacted people can be found by examining events in the user's calendar and determining who is likely to attend them, perhaps based on email exchanges or former events' participant lists. Of course, uncertainty has to be maintained along the entire process, and provenance information is needed to explain query results to the user (e.g., indicate which meetings and trips are relevant to each person of the output). Knowledge about services, their data models, their costs, need either to be provided by the system designer, or to be automatically learned from interaction with these services, as in [68].

One motivation for that choice is that Pims concentrate many of the problems we intend to investigate: heterogeneity (various sources, each with a different structure), massive distribution (information spread out over the Web, in numerous sources), rapid evolution (new data regularly added), intensionality (knowledge from Wikidata, OpenStreetMap...), confidentiality and security (mostly private data), and uncertainty (very variable quality). Though the data is distributed, its size is relatively modest; other applications may be considered for works focusing on processing data at large scale, which is a potential research direction within Valda, though not our main focus. Another strong motivation for the choice of Pims as application domain is the importance of this application from a societal viewpoint.

A Pims is essentially a system built on top of a user's *personal knowledge base*; such knowledge bases are reminiscent of those found in the Semantic Web, e.g., linked open data. Some issues, such as ontology alignment [71] exist in both scenarios. However, there are some fundamental differences in building personal knowledge bases vs collecting information from the Semantic Web: first, the scope is quite smaller, as one is only interested in knowledge related to a given individual; second, a small proportion of the data is already present in the form of semantic information, most needs to be extracted and annotated through appropriate wrappers and enrichers; third, though the linked open data is meant to be read-only, the only update possible to a user being adding new triples, a personal knowledge base is very much something that a user needs to be able to edit, and propagating updates from the knowledge base to original data sources is a challenge in itself.

4.2. Web Data

The choice of Pims is not exclusive. We also consider other application areas as well. In particular, we have worked in the past and have a strong expertise on Web data [38] in a broad sense: semi-structured, structured, or unstructured content extracted from Web databases [68]; knowledge bases from the Semantic Web [71]; social networks [66]; Web archives and Web crawls [54]; Web applications and deep Web databases [47]; crowdsourcing platforms [41]. We intend to continue using Web data as a natural application domain for the research within Valda when relevant. For instance [45], deep Web databases are a natural application scenario for intensional data management issues: determining if a deep Web database contains some information requires optimizing the number of costly requests to that database.

A common aspect of both personal information and Web data is that their exploitation raises ethical considerations. Thus, a user needs to remain fully in control of the usage that is made of her personal information; a search engine or recommender system that ranks Web content for display to a specific user needs to do so in an unbiased, justifiable, manner. These ethical constraints sometimes forbid some technically solutions that may be technically useful, such as sharing a model learned from the personal data of a user to another user, or using blackboxes to rank query result. We fully intend to consider these ethical considerations within Valda. One of the main goals of a Pims is indeed to empower the user with a full control on the use of this data.

5. Highlights of the Year

5.1. Highlights of the Year

Leonid Libkin, formerly Professor at the University of Edinburgh, was recruited as a senior member of the group in 2019, first (from September to November 2019), with a *Chaire d'Excellence* from FSMP (Fédération des Sciences Mathématiques de Paris), and then as a Professor at ENS.

5.1.1. Awards

Mikaël Monet received the 2019 PhD award of the French database community (BDA) for his PhD prepared within Valda and defended in 2018.

6. New Software and Platforms

6.1. ProvSQL

KEYWORDS: Databases - Provenance - Probability

FUNCTIONAL DESCRIPTION: The goal of the ProvSQL project is to add support for (m-)semiring provenance and uncertainty management to PostgreSQL databases, in the form of a PostgreSQL extension/module/plugin.

NEWS OF THE YEAR: Miscellaneous enhancements and bug fixes. Addition of a tutorial.

- Participants: Pierre Senellart and Yann Ramusat
- Contact: Pierre Senellart
- Publications: Provenance and Probabilities in Relational Databases: From Theory to Practice - ProvSQL: Provenance and Probability Management in PostgreSQL
- URL: https://github.com/PierreSenellart/provsql

6.2. apxproof

KEYWORD: LaTeX

FUNCTIONAL DESCRIPTION: apxproof is a LaTeX package facilitating the typesetting of research articles with proofs in appendix, a common practice in database theory and theoretical computer science in general. The appendix material is written in the LaTeX code along with the main text which it naturally complements, and it is automatically deferred. The package can automatically send proofs to the appendix, can repeat in the appendix the theorem environments stated in the main text, can section the appendix automatically based on the sectioning of the main text, and supports a separate bibliography for the appendix material.

RELEASE FUNCTIONAL DESCRIPTION: Fix formatting of theorems (and proof sketches) to be faithful to the way they are formatted in the base document class (this will change some difference in the appearance of documents typset with earlier versions of apxproof), Configurable mainbodyrepeatedtheorem command to add some styling to repeated theorems, Allow using apxproof without bibunits (e.g., for biblatex compatibility), Restore predefined theorem counters, allowing more robust use of apxproof when the base document class predefines theorems.

NEWS OF THE YEAR: Major 1.2.0 release with a much more faithful rendering of theorems compared to the original base classes, bug fixes, compatibility enhancements (in particular, with respect to the use of biblatex or of fancyvrb).

- Participant: Pierre Senellart
- Contact: Pierre Senellart
- URL: https://github.com/PierreSenellart/apxproof

7. New Results

7.1. Foundations of data management

We obtained a number of results on the foundations of data management, i.e., in database theory.

We worked on **knowledge bases**. In our work a knowledge base consists of an incomplete database together with a set of existential rules. We investigated the problem of query answering: computing the answers that are logically entailed from the knowledge base. This brings to light the fundamental chase tool, and its different variants that have been proposed in the literature. We studied the problem of chase termination, which has applications beyond query answering, and studied its complexity for restricted but useful classes of existential rules [27].

We worked on **data integration**. In our scenario a user can access data sitting in multiple sources by means of queries over a global schema, related to the sources via mappings. Data sources often contain sensitive information, and thus an analysis is needed to verify that a schema satisfies a privacy policy, given as a set of queries whose answers should not be accessible to users. We show that source constraints can have a dramatic impact on disclosure analysis [22]. Another work related to data integration is [16], where we connect the problem of answering queries under limited accesses (e.g., using Web forms) to two foundational issues: containment of Monadic datalog (MDL) programs, and containment problems involving regular tree languages. In particular, we establish a 2EXPTIME lower bound on the problem of containment of a MDL program into a conjunctive query, resolving an open problem from the early 1990s.

We also considered some other foundational topics, further from core database topics. In [18], we establish bounds on the height of maximal finite towers (a *tower* is a sequence of words alternating between two languages in such a way that every word is a subsequence of the following word) between two regular languages. In [17], we present an online $O(\sigma|y|)$-time algorithm for finding approximate occurrences of a word x within a word y, where σ is the alphabet size.

Note that two other works in this theme will be described in the 2020 activity report, as they are published in 2020 conferences [25], [26].

7.2. Uncertainty and provenance of data

We have a strong focus on the uncertainty and provenance in databases. See [20] for a high-level introduction to the area.

In [15], we investigate the use of knowledge compilation, i.e., obtaining compact circuit-based representations of functions, for (Boolean) provenance. Some width parameters of the circuit, such as bounded treewidth or pathwidth, can be leveraged to convert the circuit to structured classes, e.g., deterministic structured NNFs (d-SDNNFs) or OBDDs. In [14], we investigate parameterizations of both database instances and queries that make query evaluation fixed-parameter tractable in combined complexity. We show that clique-frontier-guarded Datalog with stratified negation (CFG-Datalog) enjoys bilinear-time evaluation on structures of bounded treewidth for programs of bounded rule size. Such programs capture in particular conjunctive queries with simplicial decompositions of bounded width, guarded negation fragment queries of bounded CQ-rank, or two-way regular path queries. Our result is shown by translating to alternating two-way automata, whose semantics is defined via cyclic provenance circuits (cycluits) that can be tractably evaluated.

In previous work [39], [40]. we have shown that the only restrictions to database instances that make probabilistic query evaluation tractable for a large class of queries is that of having a small treewidth. In [28], [32], we provide the first large-scale experimental study of treewidth and tree decompositions of real-world database instances (25 datasets from 8 different domains, with sizes ranging from a few thousand to a few million vertices). The goal is to determine which data, if any, has reasonably low treewidth. We also show that, even when treewidth is high, using partial tree decompositions can result in data structures that can assist algorithms.

To conclude on provenance management, in [23], [24], after investigating the complexity of satisfiability and query answering for attributed DL-LiteR ontologies, we propose a new semantics, based on provenance semirings, for integrating provenance information with query answering. Finally, we establish complexity results for satisfiability and query answering under this semantics.

We also consider **other notions of incompleteness**, such as in [13], where we study the complexity of query evaluation for databases whose relations are partially ordered; the problem commonly arises when combining or transforming ordered data from multiple sources. We focus on queries in a useful fragment of SQL, namely positive relational algebra with aggregates, whose bag semantics we extend to the partially ordered setting. Our semantics leads to the study of two main computational problems: the possibility and certainty of query answers. We show that these problems are respectively NP-complete and coNP-complete, but identify tractable cases depending on the query operators or input partial orders.

Finally, we also consider uncertainty through another angle, that of learning in a dynamic environment, using techniques from **reinforcement learning** and the **multi-armed bandit** field.

In [19], we tackle the problem of *influence maximization*: finding influential users, or nodes, in a graph so as to maximize the spread of information. We study a highly generic version of influence maximization, one of optimizing influence campaigns by sequentially selecting “spread seeds” from a set of influencers, a small subset of the node population, under the hypothesis that, in a given campaign, previously activated nodes remain persistently active. We introduce an estimator on the influencers’ remaining potential – the expected number of nodes that can still be reached from a given influencer – and justify its strength to rapidly estimate the desired value, relying on real data gathered from Twitter. We then describe a novel algorithm, GT-UCB, relying on probabilistic upper confidence bounds on the remaining potential.

In [21], we propose a Bayesian information-geometric approach to the exploration-exploitation trade-off in stochastic multi-armed bandits. The uncertainty on reward generation and belief is represented using the manifold of joint distributions of rewards and beliefs. Accumulated information is summarised by the barycentre of joint distributions, the pseudobelief-reward. While the pseudobelief-reward facilitates information accumulation through exploration, another mechanism is needed to increase exploitation by gradually focusing on higher rewards, the pseudobelief-focal-reward. Our resulting algorithm, BelMan, alternates between projection of the pseudobelief-focal-reward onto belief-reward distributions to choose the arm to play, and projection of the updated belief-reward distributions onto the pseudobelief-focal-reward.

In [29], we consider another form of bandits, *linear bandits*, in which the available actions correspond to arbitrary context vectors whose associated rewards follow a non-stationary linear regression model. In this setting, the unknown regression parameter is al- lowed to vary in time. To address this problem, we propose D-LinUCB, a novel optimistic algorithm based on discounted linear regression, where exponential weights are used to smoothly forget the past.

7.3. Web data management

We finally describe research more oriented towards applications.

The PhD of Karima Rafes [11] dealt with **semantic knowledge bases** and their applications to the management of scientific data, through the development of the LinkedWiki platform. Another practical work on semantic knowledge bases is [30], where we show how the edit history of of a knowledge base can help correct constraint violations.

Finally, we investigate **transparency and bias** in data management and artificial intelligence. [12] presents to the data management community the challenges raised by new regulatory frameworks in this area. In [31], we discuss the possibility for artificial intelligence systems to be used in the practice of law.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

Juliette Achddou's PhD research is set up as a CIFRE contract and supervision agreement between her employer, the Numberly company, and École normale supérieure.

We are in the process of finalizing a contract with Neo4j, the leading company in the field of graph databases, to work towards the creation of a new standard for graph languages called GQL, building on Neo4j's Cypher query language. On this, we do not start from scratch. In a joint effort between the Neo4j's Cypher group and the Edinburgh database group led by Leonid Libkin, a formal specification of the core querying and update features of Cypher was produced. Starting in 2020, Libkin will chair a working group on the formal semantics of GQL. In addition to Valda, it will involve researchers from Edinburgh, Santiago, Warsaw, and other universities in Paris (Marne-la-Vallee and Paris-Diderot). The project is supported by a grant from Neo4j.

9. Partnerships and Cooperations

9.1. Regional Initiatives

The ISORE project from the Île-de-France region (6k€ grant, DIM RFSI), which started in 2019, was completed in 2020.

Leonid Libkin received funding from FSMP through his *Chaire d'Excellence*, in the fall of 2019.

Pierre Senellart is a recipient of a Chair of the PaRis Artificial Intelligence Research InstitutE, PRAIRIE, sarting in the fall of 2019.

9.2. National Initiatives

9.2.1. ANR

Valda has been part of four ANR projects in 2019:

HEADWORK (2016–2021; 38 k€ for Valda, budget managed by Inria), together with IRISA (Druid, coordinator), Inria Lille (Links & Spirals), and Inria Rennes (Sumo), and two application partners: MNHN (Cesco) and FouleFactory. The topic is workflows for crowdsourcing. See http://headwork.gforge.inria.fr/.

BioQOP (2017–2020; 66 k€ for Valda, budget managed by ENS), with Idemia (coordinator) and GREYC, on the optimization of queries for privacy-aware biometric data management. See http://bioqop.di.ens.fr/.

CQFD (2018–2022; 19 k€ for Valda, budget managed by Inria), with Inria Sophia (GraphIK, coordinator), LaBRI, LIG, Inria Saclay (Cedar), IRISA, Inria Lille (Spirals), and Télécom ParisTech, on complex ontological queries over federated and heterogeneous data. See http://www.lirmm.fr/cqfd/.

QUID (2018–2022; 49 k€ for Valda, budget managed by Inria), LIGM (coordinator), IRIF, and LaBRI, on incomplete and inconsistent data. See https://quid.labri.fr/home.html.

Camille Bourgaux is participating in the AI Chair of Meghyn Bienvenu on *INTENDED (Intelligent handling of imperfect data)* to start in 2020.

9.3. European Initiatives

9.3.1. Collaborations in European Programs, Except FP7 & H2020

A bilateral French–German ANR project, entitled *EQUUS – Efficient Query answering Under UpdateS* was accepted in 2019. It will start in 2020. It involves CNRS (CRIL, CRIStAL, IMJ), Télécom Paris, HU Berlin, and Bayreuth University, in addition to Inria Valda.

9.4. International Initiatives

9.4.1. Informal International Partners

Valda has strong collaborations with the following international groups:

Univ. Edinburgh, United Kingdom: Paolo Guagliardo, Andreas Pieris

Univ. Oxford, United Kingdom: Michael Benedikt, Dan Olteanu, and Georg Gottlob

TU Dresden, Germany: Markus Krötzsch and Sebastian Rudolph

Dortmund University, Germany: Thomas Schwentick

Free Univ. Bozen-Bolzano, Italy: Ana Ozaki

Warsaw University, Poland: Mikołaj Bojańczyk and Szymon Toruńczyk

Tel Aviv University, Israel: Daniel Deutch and Tova Milo

Drexel University, USA: Julia Stoyanovich

Univ. California San Diego, USA: Victor Vianu

Pontifical Catholic University of Chile: Marcelo Arenas, Pablo Barceló

National University of Singapore: Stéphane Bressan

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Victor Vianu, Professor at UC San Diego and former holder of an Inria international chair, spent 6 months within Valda, as a University Paris-Diderot and ENS invited professor.

Thomas Schwentick, Professor at TU Dortmund, spend 1 month within Valda in May–June.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

10.1.1.1. General Chair, Scientific Chair

- Camille Bourgaux, organizer of the yearly meeting of the national working group on *Automata, Logic, Games, and Algebra* (ALGA, GDR IM) in 2019
- Leonid Libkin was appointed general chair of PODS
- Luc Segoufin, chair of the steering committee of the conference series *Highlights of Logic, Games and Automata*
- Luc Segoufin and Pierre Senellart, co-organizers of École de Printemps en Informatique Théorique (EPIT) 2019
- Pierre Senellart, co-organizer and chief judge of the ICPC (International Collegiate Programming Contest) Southwestern Europe 2019-2020 competition

10.1.2. Scientific Events: Selection

10.1.2.1. Chair of Conference Program Committees

- Leonid Libkin, LICS 2021 (in 2019: constitution of the program committee)

10.1.2.2. Member of the Conference Program Committees

- Camille Bourgaux, IJCAI 2019, AAAI 2020, ECAI 2020
- Olivier Cappé, ALT 2019
- Leonid Libkin, IJCAI 2019, AAAI 2020, ECAI 2020, FOSSACS 2020, ICDT 2020
- Pierre Senellart, SUM 2019, PODS 2019, STACS 2020
- Michaël Thomazo, IJCAI 2019

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- Olivier Cappé, *Annals of the Institute of Statistical Mathematics*
- Leonid Libkin, *Acta Informatica*
- Leonid Libkin, *Bulletin of Symbolic Logic*
- Leonid Libkin, *Journal of Applied Logic*
- Leonid Libkin, *SN Computer Science*

10.1.3.2. Reviewer - Reviewing Activities

- Pierre Senellart, *Future Generation Computer Systems*

10.1.4. Invited Talks

- Leonid Libkin, keynote at BDA 2019
- Pierre Senellart, invited talk at Singaporean-French workshop on Artificial Intelligence (SinFra), Singapore
- Pierre Senellart, invited lecture at *Reasoning Web* summer school, Bolzano, Italy

10.1.5. Leadership within the Scientific Community

- Serge Abiteboul is a member of the French Academy of Sciences, of the Academia Europaea, of the scientific council of the Société Informatique de France, and an ACM Fellow.
- Leonid Libkin is a Fellow of the Royal Society of Edinburgh, a member of the Academia Europaea, of the UK Computing research committee, and an ACM Fellow.
- Pierre Senellart is a member of the steering committee of BDA, the French scientific community on data management.

10.1.6. Scientific Expertise

- Pierre Senellart has performed a confidential audit of a company for the French government (*direction interministérielle du numérique et du système d'information et de communication de l'État*)
- Pierre Senellart, ANR

10.1.7. Research Administration

- Olivier Cappé is a scientific deputy director of CNRS division of Information Sciences and Technologies (INS2I).
- Luc Segoufin is a member of the CNHSCT of Inria.
- Pierre Senellart is a member of the board of section 6 of the National Committee for Scientific Research.

- Pierre Senellart is deputy director of the DI ENS laboratory, joint between ENS, CNRS, and Inria.
- Pierre Senellart is a member of the board of the DIM RFSI (Réseau Francilien en Sciences Informatiques).
- Pierre Senellart was a member of the scientific council of PGMO (Programme Gaspard Monge) until mid-2019.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence: Serge Abiteboul, *Scientific reading group*, 15 heqTD, L3, École normale supérieure

Licence: Pierre Senellart, Nathan Grosshans, Michaël Thomazo, *Databases*, 74 heqTD, L3, École normale supérieure

Licence: Pierre Senellart, *Algorithms*, 18 heqTD, L3, École normale supérieure

Licence: Nathan Grosshans, *Formal languages*, 22 heqTD, L3, Télécom ParisTech

Master: Pierre Senellart, *Web data management*, 36 heqTD, M2, MPRI

Pierre Senellart has various teaching responsibilities (L3 internships, M1 projects, M2 administration, entrance competition) at ENS. Nathan Grosshans is the secretary of the entrance competition at ENS for computer science. Most members of the group are also involved in tutoring ENS students, advising them on their curriculum, their internships, etc. They are also occasionally involved with reviewing internship reports, supervising student projects, etc.

10.2.2. Supervision

PhD: Karima Rafes, *Le Linked Data à l'université : la plateforme LinkedWiki*, Université Paris-Saclay, 25 January 2019, Serge Abiteboul & Sarah Cohen-Boulakia

PhD in progess: Juliette Achddou, *Application of reinforcement learning strategies to the context of Real-Time Bidding*, started in September 2018, Olivier Cappé & Aurélien Garivier

PhD in progress: Julien Grange, *Graph properties: order and arithmetic in predicate logics*, started in September 2017, Luc Segoufin

PhD in progress: Yann Ramusat, *Provenance-based routing in probabilistic graphs*, started in September 2018, Silviu Maniu & Pierre Senellart

PhD in progess: Yoan Russac, *Sequential methods for robust decision making*, started in December 2018, Olivier Cappé

10.2.3. Juries

- HdR Paolo Papotti, April 2019, Université de Nice – Sophia-Antipolis, Pierre Senellart (reviewer)
- PhD Ugo Comignani, September 2019, Université Claude Bernard Lyon 1, Pierre Senellart (reviewer)

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

Serge Abiteboul is the president of the strategic committee of the Blaise Pascal foundation for scientific mediation.

Pierre SENELLART is a research fellow within the CERRE (Centre on Regulation in Europe), a European think tank that produces policy papers and organize events about the regulation of network industries. He contributes in particular to reflections on the use of artificial intelligence techniques and on the interoperability of software platforms.

10.3.2. Articles and contents

Serge Abiteboul writes regular columns on popularization of computer science in La Recherche and Le Monde (Économie). He is a founding editor of the *binaire* blog for popularizing computer science. See https://www.lemonde.fr/blog/binaire/.

11. Bibliography

Major publications by the team in recent years

[1] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Explanations and Transparency in Collaborative Workflows*, in "PODS 2018 - 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles Of Database Systems", Houston, Texas, United States, June 2018, https://hal.inria.fr/hal-01744978

[2] A. AMARILLI, F. CAPELLI, M. MONET, P. SENELLART. *Connecting Knowledge Compilation Classes and Width Parameters*, in "Theory of Computing Systems", June 2019, https://arxiv.org/abs/1811.02944 [*DOI :* 10.1007/s00224-019-09930-2], https://hal.inria.fr/hal-02163749

[3] C. BOURGAUX, A. OZAKI. *Querying Attributed DL-Lite Ontologies Using Provenance Semirings*, in "Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)", Honolulu, United States, January 2019, https://hal.inria.fr/hal-02109645

[4] F. JACQUEMARD, L. SEGOUFIN, J. DIMINO. *FO2(<, +1, ~) on data trees, data tree automata and branching vector addition systems*, in "Logical Methods in Computer Science", 2016, vol. 12, n° 2, https://doi.org/10.2168/LMCS-12(2:3)2016

[5] P. LAGRÉE, O. CAPPÉ, B. CAUTIS, S. MANIU. *Algorithms for Online Influencer Marketing*, in "ACM Transactions on Knowledge Discovery from Data (TKDD)", January 2019, vol. 13, n° 1, pp. 1-30 [*DOI :* 10.1145/3274670], https://hal.inria.fr/hal-01478788

[6] M. LECLÈRE, M.-L. MUGNIER, M. THOMAZO, F. ULLIANA. *A Single Approach to Decide Chase Termination on Linear Existential Rules*, in "ICDT 2019 - International Conference on Database Theory", Lisbonne, Portugal, 2019 [*DOI :* 10.4230/LIPIcs.ICDT.2019.15], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02148200

[7] S. MANIU, R. CHENG, P. SENELLART. *An Indexing Framework for Queries on Probabilistic Graphs*, in "ACM Trans. Datab. Syst", 2017, https://hal.inria.fr/hal-01437580

[8] Y. RUSSAC, C. VERNADE, O. CAPPÉ. *Weighted Linear Bandits for Non-Stationary Environments*, in "NeurIPS 2019 - 33rd Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, https://arxiv.org/abs/1909.09146 , https://hal.inria.fr/hal-02291460

[9] N. SCHWEIKARDT, L. SEGOUFIN, A. VIGNY. *Enumeration for FO Queries over Nowhere Dense Graphs*, in "PODS 2018 - Principles Of Database Systems", Houston, United States, June 2018, https://hal.inria.fr/hal-01895786

[10] P. SENELLART, L. JACHIET, S. MANIU, Y. RAMUSAT. *ProvSQL: Provenance and Probability Management in PostgreSQL*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2018, vol. 11, n° 12, pp. 2034-2037 [*DOI :* 10.14778/3229863.3236253], https://hal.inria.fr/hal-01851538

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] K. RAFES. *Linked Data at university : the LinkedWiki platform*, Université Paris-Saclay, January 2019, https://tel.archives-ouvertes.fr/tel-02003672

Articles in International Peer-Reviewed Journals

[12] S. ABITEBOUL, J. STOYANOVICH. *Transparency, Fairness, Data Protection, Neutrality: Data Management Challenges in the Face of New Regulation*, in "Journal of data and information quality", 2019 [*DOI :* 10.1145/3310231], https://hal.inria.fr/hal-02066516

[13] A. AMARILLI, M. L. BA, D. DEUTCH, P. SENELLART. *Computing Possible and Certain Answers over Order-Incomplete Data*, in "Theoretical Computer Science", 2019, vol. 797, pp. 42-76, https://arxiv.org/abs/1801.06396 [*DOI :* 10.1016/J.TCS.2019.05.013], https://hal.inria.fr/hal-01891814

[14] A. AMARILLI, P. BOURHIS, M. MONET, P. SENELLART. *Evaluating Datalog via Tree Automata and Cycluits*, in "Theory of Computing Systems", 2019, vol. 63, n° 7, pp. 1620-1678, https://arxiv.org/abs/1808.04663 [*DOI :* 10.1007/S00224-018-9901-2], https://hal.inria.fr/hal-01891811

[15] A. AMARILLI, F. CAPELLI, M. MONET, P. SENELLART. *Connecting Knowledge Compilation Classes and Width Parameters*, in "Theory of Computing Systems", June 2019, https://arxiv.org/abs/1811.02944 [*DOI :* 10.1007/S00224-019-09930-2], https://hal.inria.fr/hal-02163749

[16] M. BENEDIKT, P. BOURHIS, G. GOTTLOB, P. SENELLART. *Monadic Datalog, Tree Validity, and Limited Access Containment*, in "ACM Transactions on Computational Logic", October 2019, vol. 21, n° 1, pp. 6:1-6:45 [*DOI :* 10.1145/3344514], https://hal.inria.fr/hal-02307999

[17] M. CROCHEMORE, A. HÉLIOU, G. KUCHEROV, L. MOUCHARD, S. PISSIS, Y. RAMUSAT. *Absent words in a sliding window with applications*, in "Information and Computation", September 2019, 104461 p. [*DOI :* 10.1016/J.IC.2019.104461], https://hal.archives-ouvertes.fr/hal-02414839

[18] S. HOLUB, T. MASOPUST, M. THOMAZO. *On the Height of Towers of Subsequences and Prefixes*, in "Information and Computation", April 2019 [*DOI :* 10.1016/J.IC.2019.01.004], https://hal.inria.fr/hal-02269576

[19] P. LAGRÉE, O. CAPPÉ, B. CAUTIS, S. MANIU. *Algorithms for Online Influencer Marketing*, in "ACM Transactions on Knowledge Discovery from Data (TKDD)", January 2019, vol. 13, n° 1, pp. 1-30 [*DOI :* 10.1145/3274670], https://hal.inria.fr/hal-01478788

Invited Conferences

[20] P. SENELLART. *Provenance in Databases: Principles and Applications*, in "RW 2019 : Reasoning Web Summer School", Bolzano, Italy, September 2019, pp. 104-109 [*DOI :* 10.1007/978-3-030-31423-1_3], https://hal.inria.fr/hal-02293688

International Conferences with Proceedings

[21] D. BASU, P. SENELLART, S. BRESSAN. *BelMan: An Information-Geometric Approach to Stochastic Bandits*, in "ECML/PKDD - The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Würzburg, Germany, September 2019, https://hal.inria.fr/hal-02195539

[22] M. BENEDIKT, P. BOURHIS, L. JACHIET, M. THOMAZO. *Reasoning about disclosure in data integration in the presence of source constraints*, in "IJCAI 2019 - 28th International Joint Conference on Artificial Intelligence", Macao, China, August 2019, https://arxiv.org/abs/1906.00624 , https://hal.inria.fr/hal-02145369

[23] C. BOURGAUX, A. OZAKI. *Querying Attributed DL-Lite Ontologies Using Provenance Semirings (Extended Abstract)*, in "DL 2019 - 32nd International Workshop on Description Logics", Oslo, Norway, June 2019, https://hal.inria.fr/hal-02152064

[24] C. BOURGAUX, A. OZAKI. *Querying Attributed DL-Lite Ontologies Using Provenance Semirings*, in "Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)", Honolulu, United States, January 2019, https://hal.inria.fr/hal-02109645

[25] J. GRANGE, L. SEGOUFIN. *Order-Invariant First-Order Logic over Hollow Trees*, in "CSL 2020 - 28th annual conference of the European Association for Computer Science Logic", Barcelona, Spain, January 2020, vol. 23, pp. 1-23 [*DOI :* 10.4230/LIPICS.CSL.2020.23], https://hal.inria.fr/hal-02310749

[26] N. GROSSHANS. *The Power of Programs over Monoids in J*, in "14th International Conference on Language and Automata Theory and Applications (LATA 2020)", Milan, Italy, March 2020, https://arxiv.org/abs/1912.07992 , https://hal.archives-ouvertes.fr/hal-02414771

[27] M. LECLÈRE, M.-L. MUGNIER, M. THOMAZO, F. ULLIANA. *A Single Approach to Decide Chase Termination on Linear Existential Rules*, in "ICDT 2019 - International Conference on Database Theory", Lisbonne, Portugal, 2019 [*DOI :* 10.4230/LIPICS.ICDT.2019.15], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02148200

[28] S. MANIU, P. SENELLART, S. JOG. *An Experimental Study of the Treewidth of Real-World Graph Data*, in "ICDT 2019 – 22nd International Conference on Database Theory", Lisbon, Portugal, March 2019, 18 p. [*DOI :* 10.4230/LIPICS.ICDT.2019.12], https://hal.inria.fr/hal-02087763

[29] Y. RUSSAC, C. VERNADE, O. CAPPÉ. *Weighted Linear Bandits for Non-Stationary Environments*, in "NeurIPS 2019 - 33rd Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, https://arxiv.org/abs/1909.09146 , https://hal.inria.fr/hal-02291460

[30] T. P. TANON, C. BOURGAUX, F. M. SUCHANEK. *Learning How to Correct a Knowledge Base from the Edit History*, in "World Wide Web Conference", San Francisco, United States, Proceedings of the 2019 World Wide Web Conference (WWW '19), May 2019 [*DOI :* 10.1145/3308558.3313584], https://hal-imt.archives-ouvertes.fr/hal-02066041

Scientific Books (or Scientific Book chapters)

[31] S. ABITEBOUL, F. G'SELL. *Les algorithmes pourraient-ils remplacer les juges ?*, in "Le Big Data et le droit", Thèmes et Commentaire, Dalloz, 2019, https://hal.inria.fr/hal-02304016

Other Publications

[32] S. MANIU, P. SENELLART, S. JOG. *An Experimental Study of the Treewidth of Real-World Graph Data (Extended Version)*, April 2019, https://arxiv.org/abs/1901.06862 - Extended version of an article published in the proceedings of ICDT 2019, https://hal.inria.fr/hal-02087770

References in notes

[33] S. ABITEBOUL, B. ANDRÉ, D. KAPLAN. *Managing your digital life*, in "Commun. ACM", 2015, vol. 58, n° 5, pp. 32-35, http://doi.acm.org/10.1145/2670528

[34] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Comparing workflow specification languages: A matter of views*, in "ACM Trans. Database Syst.", 2012, vol. 37, n° 2, pp. 10:1-10:59, http://doi.acm.org/10.1145/2188349.2188352

[35] S. ABITEBOUL, P. BUNEMAN, D. SUCIU. *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann, 1999

[36] S. ABITEBOUL, L. HERR, J. VAN DEN BUSSCHE. *Temporal Versus First-Order Logic to Query Temporal Databases*, in "Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, 1996, Montreal, Canada", R. HULL (editor), ACM Press, 1996, pp. 49-57, http://doi.acm.org/10.1145/237661.237674

[37] S. ABITEBOUL, R. HULL, V. VIANU. *Foundations of Databases*, Addison-Wesley, 1995, http://webdam.inria.fr/Alice/

[38] S. ABITEBOUL, I. MANOLESCU, P. RIGAUX, M. ROUSSET, P. SENELLART. *Web Data Management*, Cambridge University Press, 2011, http://webdam.inria.fr/Jorge

[39] A. AMARILLI, P. BOURHIS, P. SENELLART. *Provenance Circuits for Trees and Treelike Instances*, in "Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part II", 2015, pp. 56-68, https://doi.org/10.1007/978-3-662-47666-6_5

[40] A. AMARILLI, P. BOURHIS, P. SENELLART. *Tractable Lineages on Treelike Instances: Limits and Extensions*, in "Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016", T. MILO, W. TAN (editors), ACM, 2016, pp. 355-370, http://doi.acm.org/10.1145/2902251.2902301

[41] Y. AMSTERDAMER, Y. GROSSMAN, T. MILO, P. SENELLART. *CrowdMiner: Mining association rules from the crowd*, in "PVLDB", 2013, vol. 6, n° 12, pp. 1250-1253, http://www.vldb.org/pvldb/vol6/p1250-amsterdamer.pdf

[42] P. B. BAEZA. *Querying graph databases*, in "Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013", R. HULL, W. FAN (editors), ACM, 2013, pp. 175-188, http://doi.acm.org/10.1145/2463664.2465216

[43] D. BARBARÁ, H. GARCIA-MOLINA, D. PORTER. *The Management of Probabilistic Data*, in "IEEE Trans. Knowl. Data Eng.", 1992, vol. 4, n° 5, pp. 487-502, https://doi.org/10.1109/69.166990

[44] D. BASU, Q. LIN, W. CHEN, H. T. VO, Z. YUAN, P. SENELLART, S. BRESSAN. *Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning*, in "T. Large-Scale Data- and Knowledge-Centered Systems", 2016, vol. 28, pp. 96-132, https://doi.org/10.1007/978-3-662-53455-7_5

[45] M. BENEDIKT, G. GOTTLOB, P. SENELLART. *Determining relevance of accesses at runtime*, in "Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011, June 12-16, 2011, Athens, Greece", M. LENZERINI, T. SCHWENTICK (editors), ACM, 2011, pp. 211-222, http://doi.acm.org/10.1145/1989284.1989309

[46] M. BENEDIKT, P. SENELLART. *Databases*, in "Computer Science, The Hardware, Software and Heart of It", Springer, 2011, pp. 169-229, https://doi.org/10.1007/978-1-4614-1168-0_10

[47] M. BIENVENU, D. DEUTCH, D. MARTINENGHI, P. SENELLART, F. M. SUCHANEK. *Dealing with the Deep Web and all its Quirks*, in "Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012", M. BRAMBILLA, S. CERI, T. FURCHE, G. GOTTLOB (editors), CEUR Workshop Proceedings, CEUR-WS.org, 2012, vol. 884, pp. 21-24, http://ceur-ws.org/Vol-884/VLDS2012_p21_Bienvenu.pdf

[48] M. BOJAŃCZYK, L. SEGOUFIN, S. TORUŃCZYK. *Verification of database-driven systems via amalgamation*, in "Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013", R. HULL, W. FAN (editors), ACM, 2013, pp. 63-74, http://doi.acm.org/10.1145/2463664.2465228

[49] P. BUNEMAN, S. KHANNA, W.-C. TAN. *Why and Where: A Characterization of Data Provenance*, in "Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings.", J. VAN DEN BUSSCHE, V. VIANU (editors), Lecture Notes in Computer Science, Springer, 2001, vol. 1973, pp. 316-330, https://doi.org/10.1007/3-540-44503-X_20

[50] B. COURCELLE. *The Monadic Second-Order Logic of Graphs. I. Recognizable Sets of Finite Graphs*, in "Inf. Comput.", 1990, vol. 85, n^o 1, pp. 12-75, https://doi.org/10.1016/0890-5401(90)90043-H

[51] N. N. DALVI, D. SUCIU. *The dichotomy of probabilistic inference for unions of conjunctive queries*, in "J. ACM", 2012, vol. 59, n^o 6, pp. 30:1-30:87, http://doi.acm.org/10.1145/2395116.2395119

[52] A. DESHPANDE, Z. G. IVES, V. RAMAN. *Adaptive Query Processing*, in "Foundations and Trends in Databases", 2007, vol. 1, n^o 1, pp. 1-140, https://doi.org/10.1561/1900000001

[53] P. DONMEZ, J. G. CARBONELL. *Proactive learning: cost-sensitive active learning with multiple imperfect oracles*, in "Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008", J. G. SHANAHAN, S. AMER-YAHIA, I. MANOLESCU, Y. ZHANG, D. A. EVANS, A. KOLCZ, K. CHOI, A. CHOWDHURY (editors), ACM, 2008, pp. 619-628, http://doi.acm.org/10.1145/1458082.1458165

[54] M. FAHEEM, P. SENELLART. *Adaptive Web Crawling Through Structure-Based Link Classification*, in "Digital Libraries: Providing Quality Information - 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015, Proceedings", R. B. ALLEN, J. HUNTER, M. L. ZENG (editors), Lecture Notes in Computer Science, Springer, 2015, vol. 9469, pp. 39-51, https://doi.org/10.1007/978-3-319-27974-9_5

[55] L. GETOOR. *Introduction to statistical relational learning*, MIT Press, 2007

[56] G. GOURITEN, S. MANIU, P. SENELLART. *Scalable, generic, and adaptive systems for focused crawling*, in "25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014", L. FERRES, G. ROSSI, V. A. F. ALMEIDA, E. HERDER (editors), ACM, 2014, pp. 35-45, http://doi.acm.org/10.1145/2631775.2631795

[57] T. J. GREEN, G. KARVOUNARAKIS, V. TANNEN. *Provenance semirings*, in "Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China", L. LIBKIN (editor), ACM, 2007, pp. 31-40, http://doi.acm.org/10.1145/1265530.1265535

[58] T. J. GREEN, V. TANNEN. *Models for Incomplete and Probabilistic Information*, in "IEEE Data Eng. Bull.", 2006, vol. 29, n^{o} 1, pp. 17-24, http://sites.computer.org/debull/A06mar/green.ps

[59] A. Y. HALEVY. *Answering queries using views: A survey*, in "VLDB J.", 2001, vol. 10, n^{o} 4, pp. 270-294, https://doi.org/10.1007/s007780100054

[60] M. A. HEARST, S. T. DUMAIS, E. OSUNA, J. PLATT, B. SCHOLKOPF. *Support vector machines*, in "IEEE Intelligent Systems", 1998, vol. 13, n^{o} 4, pp. 18-28, https://doi.org/10.1109/5254.708428

[61] T. IMIELINSKI, W. LIPSKI JR.. *Incomplete Information in Relational Databases*, in "J. ACM", 1984, vol. 31, n^{o} 4, pp. 761-791, http://doi.acm.org/10.1145/1634.1886

[62] B. KIMELFELD, P. SENELLART. *Probabilistic XML: Models and Complexity*, in "Advances in Probabilistic Databases for Uncertain Information Management", Z. MA, L. YAN (editors), Studies in Fuzziness and Soft Computing, Springer, 2013, vol. 304, pp. 39-66, https://doi.org/10.1007/978-3-642-37509-5_3

[63] A. C. KLUG. *Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions*, in "J. ACM", 1982, vol. 29, n^{o} 3, pp. 699-717, http://doi.acm.org/10.1145/322326.322332

[64] D. KOSSMANN. *The State of the art in distributed query processing*, in "ACM Comput. Surv.", 2000, vol. 32, n^{o} 4, pp. 422-469, http://doi.acm.org/10.1145/371578.371598

[65] J. D. LAFFERTY, A. MCCALLUM, F. C. N. PEREIRA. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, in "Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001", C. E. BRODLEY, A. P. DANYLUK (editors), Morgan Kaufmann, 2001, pp. 282-289

[66] S. LEI, S. MANIU, L. MO, R. CHENG, P. SENELLART. *Online Influence Maximization*, in "Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015", 2015, pp. 645-654, http://doi.acm.org/10.1145/2783258.2783271

[67] F. NEVEN. *Automata Theory for XML Researchers*, in "SIGMOD Record", 2002, vol. 31, n^{o} 3, pp. 39-46, http://doi.acm.org/10.1145/601858.601869

[68] P. SENELLART, A. MITTAL, D. MUSCHICK, R. GILLERON, M. TOMMASI. *Automatic wrapper induction from hidden-web sources with domain knowledge*, in "10th ACM International Workshop on Web Information

and Data Management (WIDM 2008), Napa Valley, California, USA, October 30, 2008", C. Y. CHAN, N. POLYZOTIS (editors), ACM, 2008, pp. 9-16, http://doi.acm.org/10.1145/1458502.1458505

[69] B. SETTLES, M. CRAVEN, L. FRIEDLAND. *Active learning with real annotation costs*, in "NIPS 2008 Workshop on Cost-Sensitive Learning", 2008, http://burrsettles.com/pub/settles.nips08ws.pdf

[70] B. SETTLES. *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2012, https://doi.org/10.2200/S00429ED1V01Y201207AIM018

[71] F. M. SUCHANEK, S. ABITEBOUL, P. SENELLART. *PARIS: Probabilistic Alignment of Relations, Instances, and Schema*, in "PVLDB", 2011, vol. 5, n^{o} 3, pp. 157-168, http://www.vldb.org/pvldb/vol5/p157_fabianmsuchanek_vldb2012.pdf

[72] D. SUCIU, D. OLTEANU, C. RÉ, C. KOCH. *Probabilistic Databases*, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2011, https://doi.org/10.2200/S00362ED1V01Y201105DTM016

[73] R. S. SUTTON, A. G. BARTO. *Reinforcement learning - an introduction*, Adaptive computation and machine learning, MIT Press, 1998, http://www.worldcat.org/oclc/37293240

[74] M. Y. VARDI. *The Complexity of Relational Query Languages (Extended Abstract)*, in "Proceedings of the 14th Annual ACM Symposium on Theory of Computing, May 5-7, 1982, San Francisco, California, USA", H. R. LEWIS, B. B. SIMONS, W. A. BURKHARD, L. H. LANDWEBER (editors), ACM, 1982, pp. 137-146, http://doi.acm.org/10.1145/800070.802186

[75] K. ZHOU, M. LALMAS, T. SAKAI, R. CUMMINS, J. M. JOSE. *On the reliability and intuitiveness of aggregated search metrics*, in "22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013", Q. HE, A. IYENGAR, W. NEJDL, J. PEI, R. RASTOGI (editors), ACM, 2013, pp. 689-698, http://doi.acm.org/10.1145/2505515.2505691

[76] M. T. ÖZSU, P. VALDURIEZ. *Principles of Distributed Database Systems, Third Edition*, Springer, 2011, https://doi.org/10.1007/978-1-4419-8834-8